Available online at www.sciencedirect.com
ScienceDirect
Annals of Pure and Applied Logic 151 (2008) 62–67
ANNALS OF PURE AND APPLIED LOGIC
www.elsevier.com/locate/apal

A remark on equivalent Rosser sentences

Christopher von Bülow*

University of Constance, Germany

Received 13 October 2006; received in revised form 8 October 2007; accepted 8 October 2007
Available online 26 November 2007

Communicated by S.N. Artemov

Abstract

An oversight in Guaspari and Solovay's "Rosser sentences" [D. Guaspari, R.M. Solovay, Rosser sentences, Annals of Mathematical Logic 16 (1) (1979) 81–99] is pointed out and emended. It concerns the premisses of their proof that there are standard proof predicates all of whose Rosser sentences are provably equivalent. The result holds up, but the premisses mentioned in the paper have to be strengthened somewhat.

Keywords: Proof predicate; Rosser sentence; Provability logic

1. Introduction

In 1979 David Guaspari and Robert Solovay published a joint paper titled "Rosser sentences" [1].[1] In this paper, they introduce formal systems of propositional modal logic which capture the provability logic of Rosser sentences, i.e., of sentences in the language $\mathcal{L}_{\mathbf{PA}}$ of Peano arithmetic which assert: "there is a disproof of me that occurs before any proof of me."

For these formal systems semantical and arithmetical completeness theorems hold, which are formulated and proved by Guaspari and Solovay in Sections 3–5 of their paper. They then go on to apply their results in Parts A and B of Section 6. Thus, in Subsection 6.A they show that there are standard proof predicates[2] *not* all of whose Rosser sentences are provably equivalent. Finally, in Subsection 6.C, they prove that there also are standard proof predicates (SPP's) all of whose Rosser sentences *are* provably equivalent.

To this purpose, they define an exotic proof predicate which utilizes a list of its own Rosser sentences, the 'Rosser list', in deciding how to behave. For this reason I call the new proof predicate "LTh" — short for "(Rosser) list theorem". The formula $\mathsf{LTh}(\mathtt{f})$ will be of the form $\exists \mathtt{l}\, \mathsf{LPF}(\mathtt{l}, \mathtt{f})$, where "$\mathsf{LPF}(\mathtt{l}, \mathtt{f})$" stands for "$\mathtt{l}$ is a 'list proof'

* Corresponding address: Universität Konstanz, Philosophische Fakultät, Fach D 21, D-78457 Konstanz, Germany.
E-mail address: Christopher.von.Buelow@uni-konstanz.de.
URL: http://www.uni-konstanz.de/FuF/Philo/Philosophie/philosophie/?/88-0-Christopher-von-Buelow.htm.

[1] The present paper should be read along with [1] to supply the context not given here.

[2] The concept of a standard proof predicate is a generalization of Gödel's $\mathsf{Bew}(\mathtt{f})$, which says: "there is a **PA**-proof the last line of which is $\mathtt{f}$"; see Definition 1.

0168-0072/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.apal.2007.10.002

for $\mathtt{f}$".[3] The formula $\mathsf{LPF}(\mathtt{l},\mathtt{f})$ is a pterm (pseudo-term) with respect to $\mathtt{f}$ [2, p. 24], i.e., **PA** proves that for every $\mathtt{l}$, there is one and only one $\mathtt{f}$ such that $\mathsf{LPF}(\mathtt{l},\mathtt{f})$. (The pterm $\mathsf{LPF}(\mathtt{l},\mathtt{f})$ corresponds to the recursive function f in [1, pp. 97–98].)

In defining $\mathsf{LPF}(\mathtt{l},\mathtt{f})$, Guaspari and Solovay start with an arbitrary SPP $\mathsf{Th}(\mathtt{f})$ having a certain additional property (+), given in [1, p. 97]. For example, $\mathsf{Th}(\mathtt{f})$ might be the usual proof predicate $\mathsf{Bew}(\mathtt{f})$. From $\mathsf{Th}(\mathtt{f})$ we obtain a formula $\mathsf{ThNum}(\mathtt{m},\mathtt{f})$ which describes a numeration of the **PA**-theorems w.r.t. $\mathsf{Th}(\mathtt{f})$. (In the case of $\mathsf{Bew}(\mathtt{f})$ this could be a numeration replicating the sequence given by the size of the Gödel numbers of corresponding proofs.) Since $\mathsf{ThNum}(\mathtt{m},\mathtt{f})$ is a pterm, we can adopt a convention from [2, p. 24] and write "$\mathsf{thNum}(\mathtt{m})$" to stand for the $\mathtt{f}$ with $\mathsf{ThNum}(\mathtt{m},\mathtt{f})$, as if "$\mathsf{thNum}$" were a new one-place function letter. (This pterm corresponds to the recursive function g in [1, pp. 97–98].)

The new proof predicate $\mathsf{LTh}(\mathtt{f})$ works according to the following algorithm[4]: In the beginning we let LTh output whatever thNum outputs; that is, we let $\mathsf{LPF}\big(0,\mathsf{thNum}(0)\big)$, $\mathsf{LPF}\big(1,\mathsf{thNum}(1)\big)$, and so forth. While this is going on we also check each formula that is output whether it is by any chance one which says, for some sentence ρ:

ρ is a Rosser sentence with respect to $\mathsf{LTh}(\mathtt{f})$,

i.e., whether it is of the form,

$$\rho \;\leftrightarrow\; \exists \mathtt{y}\Big[\mathsf{LPF}\big(\mathtt{y},\ulcorner\neg\rho\urcorner\big) \;\wedge\; (\forall \mathtt{z}\leq \mathtt{y})\,\neg\mathsf{LPF}\big(\mathtt{z},\ulcorner\rho\urcorner\big)\Big].$$

If so, we add ρ to the Rosser list — or at least we do so ordinarily. However, we don't want the list to contain sentences ρ_1, ρ_2 for which $\rho_2 = \neg\rho_1$, because then we would have to keep track of which of them comes earlier in the list. In order to avoid this, we *refrain* from adding the Rosser sentence ρ to the Rosser list if the list already includes either $\neg\rho$ or some ρ' such that $\rho = \neg\rho'$.

While LTh outputs Th-theorems and, slowly, the Rosser list is being built up, we also watch whether one of these Th-theorems is (i) a Rosser sentence w.r.t. $\mathsf{LTh}(\mathtt{f})$, or (ii) the negation of one. (Actually, we merely look whether it is one of the sentences on the Rosser list in its current state, or the negation of such a sentence.) When this happens, Guaspari and Solovay say: 'the bell rings.'

In case (i), Th outputs an LTh-Rosser sentence. We let LTh reproduce this behavior by making LTh immediately output the whole Rosser list and, afterwards, all sentences of $\mathcal{L}_{\mathbf{PA}}$ in some arbitrary order.[5] Assuming that LTh hasn't yet output the *negation* of any of the sentences on the list, all list members are thus output ('proved') by LTh before their negations are. But as Rosser sentences they all assert the converse. Thus they are all false, and therefore equivalent.

In case (ii), i.e., if Th outputs the negation of a Rosser sentence, we make LTh immediately output the negations of all the sentences on the list and then (as before) all sentences of $\mathcal{L}_{\mathbf{PA}}$. Assuming that LTh hasn't yet output any of the Rosser-listed sentences themselves, their negations are thus being 'proved' before they themselves are. This is just what they say, and so they are all true and hence equivalent.

This sketch is of course far from sufficient as a *proof* that all Rosser sentences w.r.t. $\mathsf{LTh}(\mathtt{f})$ are provably equivalent; it is merely intended to provide some conception of the functioning of the new proof predicate $\mathsf{LTh}(\mathtt{f})$, and to hint at the direction which a proof would have to take. Here, it is enough to say that a proof is given in [1, pp. 97–98], but that the condition (+) offered in the paper is slightly too weak to support the proof presented for the central Lemma 6.3 [1, p. 98].[6] The point of this paper is to emend the condition and to show — in a little more detail than in the original paper — how Lemma 6.3 is to be proved.

2. How to prove Lemma 6.3

Here is Guaspari and Solovay's definition of a standard proof predicate [1, p. 83]:

[3] I employ typewriter font to set off object-language variables.

[4] Cf. [3, [24.1], [24.2]]. The proof predicate is obtained via the diagonal lemma from a formalization of the following description.

[5] Guaspari and Solovay let LTh also output the negations of the list members after outputting the members themselves, but that isn't necessary. An analogous point holds for case (ii), where Th outputs a *negated* Rosser sentence.

[6] I do not however have a proof for this claim. The same result as in [1] is given in Smoryński's [4, Lemma 3.7, p. 295] for the usual proof predicate, but he doesn't have a similar problem because his definition of "standard proof predicate" [4, p. 279] already includes provable equivalence to $\mathsf{Bew}(\mathtt{f})$, and thus our properties (+*) and (+′) (see p. 66).

Definition 1. $\mathsf{Th}(\mathtt{f})$ is a *standard proof predicate* (*SPP*) iff $\mathsf{Th}(\mathtt{f})$ is a Σ^0_1-numeration of the theorems of **PA** in **PA** — i.e., there is a Δ^0_1-formula $\mathsf{PF}(\mathtt{p},\mathtt{f})$ such that $\mathsf{Th}(\mathtt{f}) = \exists \mathtt{p}\, \mathsf{PF}(\mathtt{p},\mathtt{f})$ and for all $\mathcal{L}_{\mathbf{PA}}$-sentences φ,

(SPP1) $\quad \mathbb{N} \vDash \mathsf{Th}(\ulcorner\varphi\urcorner) \quad \Leftrightarrow \quad \mathbf{PA} \vdash \varphi$

holds — satisfying, for all sentences φ and ψ:

(SPP2) $\quad \mathbf{PA} \vdash \mathsf{Th}(\ulcorner\varphi \to \psi\urcorner) \wedge \mathsf{Th}(\ulcorner\varphi\urcorner) \to \mathsf{Th}(\ulcorner\psi\urcorner)$,

and for every Σ^0_1-sentence σ:

(SPP3) $\quad \mathbf{PA} \vdash \sigma \to \mathsf{Th}(\ulcorner\sigma\urcorner)$.

Every SPP $\mathsf{Th}(\mathtt{f})$ fulfills the Bernays–Löb derivability conditions, including, for all sentences φ:

(DC1) $\quad \mathbf{PA} \vdash \varphi \quad \Rightarrow \quad \mathbf{PA} \vdash \mathsf{Th}(\ulcorner\varphi\urcorner)$.[7]

The extra property (+) for SPP's $\mathsf{Th}(\mathtt{f})$, as given in [1, p. 97], is the following:

(+) **PA** proves: "$\{\mathtt{f} \mid \mathsf{Th}(\mathtt{f})\}$ is closed under tautological consequence and contains all true Σ^0_1-sentences",

where "true" is formalized by using the 'usual' truth predicate for Σ^0_1-formulas.

To make this more explicit, let $\mathsf{Taut}(\mathtt{f})$ be a Δ^0_1-formula expressing that $\mathtt{f}$ is (the Gödel number of) a tautological $\mathcal{L}_{\mathbf{PA}}$-sentence, and let $\Sigma\mathsf{True}(\mathtt{f})$ be Σ^0_1, saying: "$\mathtt{f}$ is a true Σ^0_1-sentence", where a sentence's being true is analyzed as its being satisfied by some variable assignment.[8] Furthermore, let $\mathsf{Cond}(\mathtt{f},\mathtt{g},\mathtt{c})$ be a Δ^0_1-pterm w.r.t. $\mathtt{c}$ which characterizes $\mathtt{c}$ as the material conditional of the formulas $\mathtt{f}$ and $\mathtt{g}$. For this $\mathtt{c}$ we subsequently write, not "$\mathsf{cond}(\mathtt{f},\mathtt{g})$", but, more suggestively, "$\mathtt{f} \ominus \mathtt{g}$". Then (+) consists in

$(+_1)$ $\quad \mathbf{PA} \vdash \forall \mathtt{f}\, \big[\mathsf{Taut}(\mathtt{f}) \to \mathsf{Th}(\mathtt{f})\big]$,

$(+_2)$ $\quad \mathbf{PA} \vdash \forall \mathtt{f},\mathtt{g}\, \Big[\mathsf{Th}(\mathtt{f} \ominus \mathtt{g}) \wedge \mathsf{Th}(\mathtt{f}) \to \mathsf{Th}(\mathtt{g})\Big]$,

$(+_3)$ $\quad \mathbf{PA} \vdash \forall \mathtt{f}\, \big[\Sigma\mathsf{True}(\mathtt{f}) \to \mathsf{Th}(\mathtt{f})\big]$.

Guaspari and Solovay's proof of Lemma 6.3, however, needs an SPP $\mathsf{Th}(\mathtt{f})$ with a little more than (+). This having been brought to his attention, Bob Solovay suggested strengthening (+) so as to include provable closure of Th under *first-order* logical consequence. The modified property (+*) then has $(+_2)$ and $(+_3)$ as before and, in place of $(+_1)$, the following:

$(+'_1)$ $\quad \mathbf{PA} \vdash \forall \mathtt{f}\, \big[\mathsf{PLTh}(\mathtt{f}) \to \mathsf{Th}(\mathtt{f})\big]$,

where $\mathsf{PLTh}(\mathtt{f})$ is Σ^0_1, saying that there is an $\mathcal{L}_{\mathbf{PA}}$-proof for sentence $\mathtt{f}$ in first-order predicate logic, that is, a proof in the language of Peano arithmetic not using any arithmetical axioms.[9]

As an anonymous referee has remarked, what we actually make use of in proving Lemma 6.3, besides $(+_2)$ and $(+_3)$, is merely a weak consequence of $(+'_1)$. — To enhance readability, I introduce some further pterms and other notation: $\mathsf{CTm}(\mathtt{t})$ is to be a Δ^0_1-formula saying that $\mathtt{t}$ is a constant term; "$\mathsf{num}(\mathtt{x})$" is used to denote (the Gödel number of) the numeral of the number $\mathtt{x}$, a constant term; "$\mathsf{subst}(\mathtt{f},\mathtt{v},\mathtt{t})$" refers to the result of substituting the term $\mathtt{t}$ for the variable $\mathtt{v}$ in the formula $\mathtt{f}$; and "$\mathsf{subst}_2(\mathtt{f};\ \mathtt{v},\mathtt{v}';\ \mathtt{t},\mathtt{t}')$" is short for

$$\mathsf{subst}\big(\mathsf{subst}(\mathtt{f},\mathtt{v},\mathtt{t}),\mathtt{v}',\mathtt{t}'\big).$$

Finally, "$\mathtt{f} \Leftrightarrow \mathtt{g}$" and "$\ominus \mathtt{f}$" are to stand for the biconditional of the formulas $\mathtt{f}$ and $\mathtt{g}$ and the negation of $\mathtt{f}$, respectively, defined in analogy to "$\mathtt{f} \ominus \mathtt{g}$". All of these pterms are Δ^0_1. — Now, as condition $(+')$ we require that over and above $(+_1)$–$(+_3)$, the following holds:

---

[7] This is because **PA** is Σ-complete and thus $\mathbb{N} \vDash \mathsf{Th}(\ulcorner\varphi\urcorner)$ implies $\mathbf{PA} \vdash \mathsf{Th}(\ulcorner\varphi\urcorner)$; cf. [5, Section 8.2].

[8] Cf. [6, Ch. 9] or [3, Sect. 23].

[9] Conditions $(+'_1)$, $(+_2)$ and (DC1) together imply $(+_3)$.

$(+_4)$ for all formulas φ and all variables $\mathtt{x}$, $\mathtt{x}'$,

$$\mathbf{PA} \vdash \forall \mathtt{t}, \mathtt{t}'\Big[\mathsf{CTm}(\mathtt{t}) \wedge \mathsf{CTm}(\mathtt{t}') \wedge \mathsf{Th}\big(\ulcorner \forall \mathtt{x}, \mathtt{x}'\, \varphi \urcorner\big) \rightarrow \mathsf{Th}\Big(\mathsf{subst}_2\big(\ulcorner\varphi\urcorner; \ulcorner \mathtt{x}\urcorner, \ulcorner \mathtt{x}'\urcorner; \mathtt{t}, \mathtt{t}'\big)\Big)\Big].^{10}$$

We prove the central lemma (cf. [3, [27.8]]) using the property $(+')$. For good measure, we also sketch a proof based on $(+_1)$–$(+_3)$ and $(+'_4)$ (see fn. 10), which is more similar to the original proof in [1].

Lemma 1. PA *proves: "if the bell rings,* $\big\{\mathtt{f} \,\big|\, \mathsf{Th}(\mathtt{f})\big\}$ *is inconsistent, i.e.,* $\mathsf{Th}\big(\ulcorner\bot\urcorner\big)$ *holds".*

Proof. Reason in **PA**. Suppose, e.g., that the bell rings in step $\mathtt{m}$ because $\mathsf{thNum}(\mathtt{m})$ itself is already on the Rosser list. Let's call it "$\mathtt{r}$" for "Rosser sentence". That $\mathtt{r}$ is on the list must be because thNum has previously output (the Gödel number of) a sentence of the form

$$\rho \leftrightarrow \exists \mathtt{y}\Big[\mathsf{LPF}\big(\mathtt{y}, \ulcorner\neg\rho\urcorner\big) \wedge (\forall \mathtt{z} \leq \mathtt{y})\, \neg\mathsf{LPF}\big(\mathtt{z}, \ulcorner\rho\urcorner\big)\Big],$$

viz., a sentence saying: "$\mathtt{r}$ is a Rosser sentence w.r.t. $\mathsf{LTh}(\mathtt{f})$." I call such a biconditional's right-hand side the "Rosser proposition for ρ". Its construction can be captured in a Δ_1^0-formula $\mathsf{RProp}(\mathtt{f}, \mathtt{p})$, a pterm w.r.t. $\mathtt{p}$, and I write "$\mathsf{rProp}(\mathtt{f})$" for such a $\mathtt{p}$. So, thNum has output $\mathtt{r} \ominus \mathsf{rProp}(\mathtt{r})$. Thus we have $\mathsf{Th}\big(\mathtt{r} \ominus \mathsf{rProp}(\mathtt{r})\big)$ and, since $\mathtt{r}$ is $\mathsf{thNum}(\mathtt{m})$, also $\mathsf{Th}(\mathtt{r})$. These two facts yield $\mathsf{Th}\big(\mathsf{rProp}(\mathtt{r})\big)$: by $(+_1)$ we have

$$\mathsf{Th}\Big(\big[\mathtt{r} \ominus \mathsf{rProp}(\mathtt{r})\big] \ominus \big[\mathtt{r} \ominus \mathsf{rProp}(\mathtt{r})\big]\Big),$$

and two applications of $(+_2)$ give us $\mathsf{Th}\big(\mathsf{rProp}(\mathtt{r})\big)$. Abbreviating the cumbersome formula $\exists \mathtt{y}\big[\mathsf{LPF}(\mathtt{y}, \mathtt{f}) \wedge (\forall \mathtt{z} \leq y)\ \neg\mathsf{LPF}(\mathtt{z}, \mathtt{g})\big]$ as "$\mathsf{Before}(\mathtt{f}, \mathtt{g})$",[11] and letting $\mathtt{n} := \ominus\mathtt{r}$, we can denote $\mathsf{rProp}(\mathtt{r})$ somewhat more transparently as "$\ulcorner\mathsf{Before}(\dot{\mathtt{n}}, \dot{\mathtt{r}})\urcorner$".[12] So, we have

$$\mathsf{Th}\big(\ulcorner\mathsf{Before}(\dot{\mathtt{n}}, \dot{\mathtt{r}})\urcorner\big), \tag{1}$$

which means that Th says: "there is an LTh-*disproof* of $\mathtt{r}$ which occurs before any LTh-*proof* of $\mathtt{r}$."

But it's just the other way round in fact: The bell rang because $\mathsf{thNum}(\mathtt{m})$ $(= \mathtt{r})$ is on the Rosser list, so, by construction, LTh at step $\mathtt{m}$ starts outputting the whole Rosser list, including $\mathtt{r}$, excluding its negation $\mathtt{n}$ (which, by construction, can't be on the list). Assume that LTh hasn't yet output $\mathtt{n}$ by step $\mathtt{m}$. (If LTh should have proved $\mathtt{n}$ before the bell rang, then it would have done so because thNum, and thus Th, endorsed $\mathtt{n}$. Since $\mathsf{Th}(\mathtt{r})$ holds as well, Th would then be inconsistent anyway.) So there is an LTh-*proof* of $\mathtt{r}$ which occurs before any LTh-*disproof* of $\mathtt{r}$:

$$\exists \mathtt{y}\Big[\mathsf{LPF}(\mathtt{y}, \mathtt{r}) \wedge (\forall \mathtt{z} \leq \mathtt{y})\, \neg\mathsf{LPF}(\mathtt{z}, \mathtt{n})\Big],$$

that is,

$$\mathsf{Before}(\mathtt{r}, \mathtt{n}).$$

---

[10] The corresponding dual requirement

$(+'_4)$ for all formulas φ and all variables $\mathtt{x}$, $\mathtt{x}'$,

$$\mathbf{PA} \vdash \forall \mathtt{t}, \mathtt{t}'\Big[\mathsf{CTm}(\mathtt{t}) \wedge \mathsf{CTm}(\mathtt{t}') \wedge \mathsf{Th}\Big(\mathsf{subst}_2\big(\ulcorner\varphi\urcorner; \ulcorner \mathtt{x}\urcorner, \ulcorner \mathtt{x}'\urcorner; \mathtt{t}, \mathtt{t}'\big)\Big) \rightarrow \mathsf{Th}\big(\ulcorner \exists \mathtt{x}, \mathtt{x}'\, \varphi \urcorner\big)\Big]$$

can do the job too. In both cases, allowing just one substitution is *not* enough. A more straightforward, but stronger, consequence of $(+'_1)$ would be

$$\mathbf{PA} \vdash \forall \mathtt{f}, \mathtt{v}, \mathtt{t}\Big[\mathsf{Fml}(\mathtt{f}) \wedge \mathsf{Var}(\mathtt{v}) \wedge \mathsf{CTm}(\mathtt{t}) \wedge \mathsf{Th}(\textcircled{\forall}\mathtt{v}\,\mathtt{f}) \rightarrow \mathsf{Th}\big(\mathsf{subst}(\mathtt{f}, \mathtt{v}, \mathtt{t})\big)\Big],$$

where $\mathsf{Fml}(\mathtt{f})$ and $\mathsf{Var}(\mathtt{v})$ are Δ_1^0-formulas saying that, respectively, $\mathtt{f}$ is a formula and $\mathtt{v}$ is a variable, and "$\textcircled{\forall}\mathtt{v}\,\mathtt{f}$" stands for the $\mathtt{v}$-universalization of $\mathtt{f}$. Again, the dual sentence works as well.

[11] This corresponds to the expression "$\mathsf{Th}_f(\mathtt{f}) \prec \mathsf{Th}_f(\mathtt{g})$" in [1].

[12] The dots above the variables indicate that this is shorthand for

$$\mathsf{subst}_2\big(\ulcorner\mathsf{Before}(\mathtt{n}, \mathtt{r})\urcorner; \ulcorner \mathtt{n}\urcorner, \ulcorner \mathtt{r}\urcorner; \mathsf{num}(\mathtt{n}), \mathsf{num}(\mathtt{r})\big),$$

in which $\mathtt{r}$ and $\mathtt{n}$ are still free.

To expose Th as inconsistent, we have to show that Th proves this too. In order to do so, we use Th's provable Σ-completeness, $(+_3)$. The expression

$$\mathsf{subst}_2\Big(\ulcorner\mathsf{Before}(\mathtt{r},\mathtt{n})\urcorner;\ \ulcorner\mathtt{r}\urcorner,\ulcorner\mathtt{n}\urcorner;\ \mathsf{num}(\mathtt{r}),\mathsf{num}(\mathtt{n})\Big)\ =\ \ulcorner\mathsf{Before}(\dot{\mathtt{r}},\dot{\mathtt{n}})\urcorner$$

is (the Gödel number of) a true Σ^0_1-sentence; hence we have

$$\Sigma\mathsf{True}\big(\ulcorner\mathsf{Before}(\dot{\mathtt{r}},\dot{\mathtt{n}})\urcorner\big),$$

which by $(+_3)$ implies

$$\mathsf{Th}\big(\ulcorner\mathsf{Before}(\dot{\mathtt{r}},\dot{\mathtt{n}})\urcorner\big). \tag{2}$$

However, $\ulcorner\mathsf{Before}(\dot{\mathtt{n}},\dot{\mathtt{r}})\urcorner$ and $\ulcorner\mathsf{Before}(\dot{\mathtt{r}},\dot{\mathtt{n}})\urcorner$ do not yet contradict each other obviously. To bring the inconsistency out into the open, we finally put to use $(+_4)$.

Let's step out of **PA** for a moment. For simple arithmetical reasons **PA** proves

$$\forall\mathtt{f},\mathtt{g}\big[\mathsf{Before}(\mathtt{f},\mathtt{g})\to\neg\mathsf{Before}(\mathtt{g},\mathtt{f})\big],$$

and thus by (DC1) also

$$\mathsf{Th}\Big(\ulcorner\forall\mathtt{f},\mathtt{g}\big[\mathsf{Before}(\mathtt{f},\mathtt{g})\to\neg\mathsf{Before}(\mathtt{g},\mathtt{f})\big]\urcorner\Big). \tag{3}$$

Now jump back inside **PA**. To obtain $\mathsf{Th}\big(\ulcorner\bot\urcorner\big)$ from (1) and (2), we have to apply the generalization in (3) to our particular 'sentences' $\mathtt{r}$ and $\mathtt{n}$. Clause $(+_4)$, but not $(+_1)$–$(+_3)$ by themselves, can give us what we need, viz., provability of the corresponding substitution instance:

$$\mathsf{Th}\big(\ulcorner\mathsf{Before}(\dot{\mathtt{r}},\dot{\mathtt{n}})\to\neg\mathsf{Before}(\dot{\mathtt{n}},\dot{\mathtt{r}})\urcorner\big).$$

An application of $(+_2)$ yields $\mathsf{Th}\big(\ulcorner\neg\mathsf{Before}(\dot{\mathtt{n}},\dot{\mathtt{r}})\urcorner\big)$, and the rest is easy: by $(+_1)$, we have

$$\mathsf{Th}\Big(\ulcorner\mathsf{Before}(\dot{\mathtt{n}},\dot{\mathtt{r}})\ \to\ \big[\neg\mathsf{Before}(\dot{\mathtt{n}},\dot{\mathtt{r}})\to\bot\big]\urcorner\Big),$$

and two more applications of $(+_2)$ finally deliver $\mathsf{Th}\big(\ulcorner\bot\urcorner\big)$.

(The proof using $(+'_4)$ runs as follows: From (1) and (2) we get (provability of) their tautological consequence,

$$\mathsf{Th}\big(\ulcorner\mathsf{Before}(\dot{\mathtt{n}},\dot{\mathtt{r}})\wedge\mathsf{Before}(\dot{\mathtt{r}},\dot{\mathtt{n}})\urcorner\big),$$

and $(+'_4)$ yields

$$\mathsf{Th}\Big(\ulcorner\exists\mathtt{f},\mathtt{g}\big[\mathsf{Before}(\mathtt{f},\mathtt{g})\wedge\mathsf{Before}(\mathtt{g},\mathtt{f})\big]\urcorner\Big).$$

Analogously to (3), however, we have

$$\mathsf{Th}\Big(\ulcorner\neg\exists\mathtt{f},\mathtt{g}\big[\mathsf{Before}(\mathtt{f},\mathtt{g})\wedge\mathsf{Before}(\mathtt{g},\mathtt{f})\big]\urcorner\Big),$$

and thus we get $\mathsf{Th}\big(\ulcorner\bot\urcorner\big)$ again.)

The other case, where the bell rings because thNum outputs the negation $\mathtt{n}$ of some Rosser sentence $\mathtt{r}$ on the list, is similar but easier. We have $\mathsf{Th}(\ominus\mathtt{r})$ and $\mathsf{Th}\big(\mathtt{r}\leftrightarrow\mathsf{rProp}(\mathtt{r})\big)$. By closure w.r.t. tautological consequence we get $\mathsf{Th}\big(\ominus\mathsf{rProp}(\mathtt{r})\big)$, i.e.,

$$\mathsf{Th}\big(\ulcorner\neg\mathsf{Before}(\dot{\mathtt{n}},\dot{\mathtt{r}})\urcorner\big).$$

But in fact $\mathsf{Before}(\mathtt{n},\mathtt{r})$ is true, because LTh outputs $\mathtt{n}=\ominus\mathtt{r}$ in or soon after step $\mathtt{m}$, when $\mathtt{r}$ hasn't yet been output (or else Th is inconsistent anyway). As a Σ^0_1-sentence, $\mathsf{Before}(\mathtt{n},\mathtt{r})$ is thus provable:

$$\mathsf{Th}\big(\ulcorner\mathsf{Before}(\dot{\mathtt{n}},\dot{\mathtt{r}})\urcorner\big).$$

Here, the contradiction in Th's theorems *is* obvious and so we don't need $(+_4)$ — nor $(+'_4)$. As before, $(+_1)$ yields

$$\mathsf{Th}\Big(\ulcorner\mathsf{Before}(\dot{\mathtt{n}},\dot{\mathtt{r}})\ \to\ \big[\neg\mathsf{Before}(\dot{\mathtt{n}},\dot{\mathtt{r}})\to\bot\big]\urcorner\Big),$$

and by twice applying $(+_2)$ we once more get $\mathsf{Th}\big(\ulcorner\bot\urcorner\big)$. □

Acknowledgements

I am grateful to Ulf Friedrichsdorf and Bob Solovay for their generous help, and to an anonymous referee for several useful suggestions.

References

[1] D. Guaspari, R.M. Solovay, Rosser sentences, Annals of Mathematical Logic 16 (1) (1979) 81–99.
[2] G. Boolos, The Logic of Provability, Cambridge University Press, Cambridge, New York, Oakleigh, 1993.
[3] C. von Bülow, Beweisbarkeitslogik: Gödel, Rosser, Solovay, in: Logische Philosophie, vol. 15, Logos, Berlin, 2006.
[4] C. Smoryński, Self-Reference and Modal Logic, in: Universitext, Springer, New York, Berlin, Heidelberg, Tokyo, 1985.
[5] J.R. Shoenfield, Mathematical Logic, in: Addison-Wesley Series in Logic, Addison-Wesley, Reading, MA, Menlo Park, CA, London, Don Mills, Ontario, 1967.
[6] R. Kaye, Models of Peano Arithmetic, Oxford University Press, Oxford, 1991.